Computational Methods for CLIP-seq Data Processing

Libertas Academica
FREEDOM TO RESEARCH

**Open Access:** Full open access to this and thousands of other papers at http://www.la-press.com.

Bioinformatics and Biology Insights

# Computational Methods for CLIP-seq Data Processing

Paula H. Reyes-Herrera[1] and Elisa Ficarra[2]

[1]Facultad de Ingeniería Electrónica y Biomédica, Universidad Antonio Nariño, Bogotá, Colombia. [2]Department of Control and Computer Engineering, Politecnico di Torino, TO, Italy.

**ABSTRACT:** RNA-binding proteins (RBPs) are at the core of post-transcriptional regulation and thus of gene expression control at the RNA level. One of the principal challenges in the field of gene expression regulation is to understand RBPs mechanism of action. As a result of recent evolution of experimental techniques, it is now possible to obtain the RNA regions recognized by RBPs on a transcriptome-wide scale. In fact, CLIP-seq protocols use the joint action of CLIP, crosslinking immunoprecipitation, and high-throughput sequencing to recover the transcriptome-wide set of interaction regions for a particular protein. Nevertheless, computational methods are necessary to process CLIP-seq experimental data and are a key to advancement in the understanding of gene regulatory mechanisms. Considering the importance of computational methods in this area, we present a review of the current status of computational approaches used and proposed for CLIP-seq data.

**KEYWORDS:** RNA-binding proteins, RBP, CLIP-based, CLIP-seq, HITS-CLIP, PAR-CLIP, RBPome, RNA–Protein, post transcriptional regulation

**CITATION:** Reyes-Herrera and Ficarra. Computational Methods for CLIP-seq Data Processing. *Bioinformatics and Biology Insights* 2014:8 199–207 doi: 10.4137/BBI.S16803.

**RECEIVED:** May 11, 2014. **RESUBMITTED:** July 29, 2014. **ACCEPTED FOR PUBLICATION:** August 1, 2014.

**ACADEMIC EDITOR:** JT Efird, Associate Editor

**TYPE:** Review

**FUNDING:** Support for this research was provided by Universidad Antonio Narifio, project grant 2012221. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.



**CORRESPONDENCE:** phreyes@gmail.com

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

## Introduction

RNA regulation is key to understanding the rules that govern gene expression regulation and epigenetic changes. RNA regulation occurs through a variety of mechanisms such as alternative splicing, alternative transcription initiation, and polyadenylation.[1,2] The systemic action of several RNA-binding proteins (RBPs) is one of the principal mechanisms of post-transcriptional gene regulation. Moreover, post-transcriptional regulation has an effect on cell function and dysfunction.[3] In fact, the correct action of each RBP and associated expression level has an impact on important processes for cell function such as cell development. At the same time RBPs' dysfunction or loss of function is associated to diseases like neurological disorders[4,5] and cancer.[6,7]

Although, the RBP role and importance are clear, and thousands of RBPs are present in eukaryotes, the mechanism of action has only been studied precisely for a few RBPs. In order to understand the RBPs' mechanism of action, it is important to identify the RBP binding sites, and from these sites the common motif. However, in humans, motifs are only known for 15% of candidate RBPs,[8] and this percentage is even lower for other organisms. Although, this field is of remarkable importance, it remains almost unexplored.

The first experimental techniques used to determine RBPs binding sites were SELEX,[9] RIP-chip,[10] and CLIP[11] (UV crosslinking and immunoprecipitation). However, these experimental techniques require a significant investment in terms of work, effort, and time. Only recently, genome-wide methods have been adapted to study RBPs' mechanism of action. In particular, CLIP-seq protocols combine the action of CLIP and next-generation sequencing (NGS) to derive a transcriptome-wide set of RBP binding sites.[12] There are

different CLIP-seq protocols; each one introduces experimental variations to improve the signal to noise ratio.

Computational methods are relevant for CLIP-seq data processing for the following reasons: CLIP-seq dataset size is significant, it is important to exploit all the information present in the experimental data, and in some cases, it is necessary to integrate other sources of information to process CLIP-seq data. Moreover, a quantitative tool to process entirely CLIP-seq data has not been developed. Instead, tools are designed to face specific CLIP-seq data processing steps. In addition, the number of CLIP-seq datasets is growing (Fig. 1). For all these reasons, computational approaches, designed specifically to deal with CLIP-seq data, are important in this field. In particular, computational methods are a key to process CLIP-seq data, facilitate the analysis, and unveil the RBP-specific roles.

This review presents the current status in computational methods designed for CLIP-seq data. Our intention is to help the reader find the most adapted tools and to motivate the readers to work on current challenges and necessities. In particular, we provide the reader with the basic background, and we present a brief overview on CLIP-seq experimental protocols and databases that contain CLIP-seq data. Moreover, we present a general computational pipeline to process CLIP-seq data and available methods at each step of the pipeline. Finally, we present future directions and current challenges.

## Background

Even though research on RBPs started in the 1970s, the interest in RBPs is concentrated mainly in the last two decades (Supplementary Fig. S1). Reviews on different topics regarding RBPs are present in the literature. We found several reviews regarding the current understanding of RNA regulation[13–17] and the RBPs role in post-transcriptional regulation.[18–20] In addition, interesting reviews are also available on the RBPs role on different organisms[21–23] and comparative studies of RBPs functionality and presence on different organisms.[8,24]

It is worth noting the reviews on the effect of RBPs on biological processes and diseases. In particular, RBP gain of function and loss of function are associated to important diseases,[6,25] like neurological diseases[4] and cancer.[7] The above mentioned reviews highlight the importance of RBP action on gene expression regulation.

Nevertheless, major advances in RBP research are marked by developments of experimental and computational techniques.[26,27] In fact, we found several reviews concerning advances in the experimental field.[28,29] In Refs. 12 and 30, the authors address the importance of *in vivo* data and the differences between *in vitro* and *in vivo* data. In Refs. 31 and 32, the authors present CLIP, crosslinking, and immunoprecipitation based approaches, and thus, the integration of high-throughput sequencing to *in vivo* experimental techniques.

On the other hand, reviews on computational approaches for RBP are mainly focused on structure prediction of Protein–RNA interactions,[33,34] as structure is key in this type of interactions.[35–39] Nonetheless, two recent reviews bring the attention to the necessity of bioinformatic approaches to process data provided by CLIP-seq protocols.[40,41]

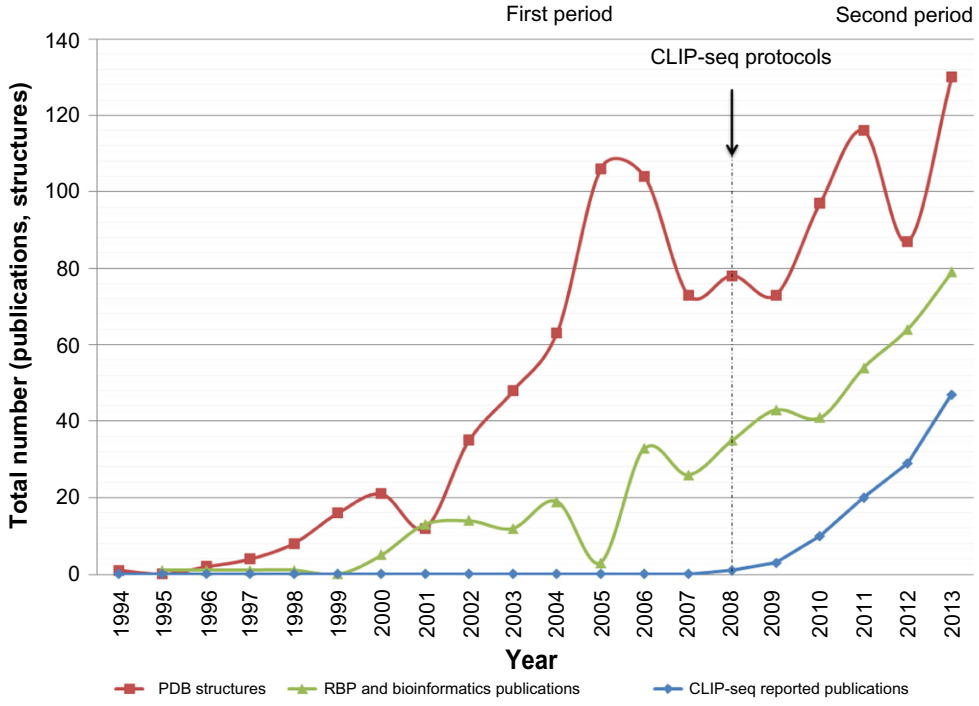



**Figure 1.** Timeline for computational research on RNA-binding proteins (RBPs). We present three indicators: red, the number of structures reported in the protein data bank; green, the number of publications of computational approaches for RBPs; and blue, the number of CLIP-seq data sets in GEO database.

In, Ref. 40 the authors present bioinformatic approaches designed for transcription factors that are frequently used to process RBP data.[40] Instead in Ref. 41, the authors emphasize the importance of modeling RNA secondary structure to recover RBP, and they conclude that RBP motif recovery is a rapidly expanding field but still in its infancy.

Considering the above, we use three indicators and associated timelines (Fig. 1) to present the focus of computational methods for RBPs. We divided the timeline into two periods, before and after the introduction of CLIP-seq protocols. In the first period, we observe a similarity between the first and second indicators trends. Instead, during the second period, we observe a similarity between the second and third indicators trends. We note that during the second period, computational proposals are essential to handle and process CLIP-seq data. Computational proposals, designed for RBPs, face two main challenges: (1) to provide insights into the RBP–RNA interaction structure and (2) to enable and facilitate CLIP-seq data processing. This review presents the computational methods designed specifically for the second challenge.

Several reviews regarding experimental advances, in particular reviews regarding CLIP-seq protocols, are currently available. On the other hand, reviews regarding computational approaches for RBPs are focused on structural prediction. However, there are no reviews focusing on computational proposals designed for CLIP-seq data, even though this is a rapidly evolving field. For these reasons, we present a review on the current status of computational proposals used and designed for CLIP-seq data. The intention of this review is to present current status, and also to motivate the readers to work on current challenges and necessities.

## CLIP-Based Experimental Data

A major step to understand the RBP role is to identify the RBP targets by locating the regions where the protein binds (also known as RNA recognition elements, RRE). The experimental field has achieved notable advances. In particular, experimental techniques used to derive RBP–RNA interactions *in vivo* are integrated with NGS technologies. As a result, it is possible to derive interaction sites on a large scale. Currently, two experimental approaches, RIP-seq and CLIP-seq, perform such integration.

RIP-seq is the combined action of RNA immunoprecipitation (RIP)[42] and RNA-seq. RIP-seq is used to recover interaction sites between RNA and specific RBPs.[32] Even though RIP-seq is simple, false positives are a major drawback. On the other hand, CLIP-seq approaches combine UV crosslinking and immunoprecipitation with NGS technologies to recover the interaction sites between RBP and RNAs. The use of UV crosslinking makes it possible to obtain reliable sites with a higher level of resolution compared to RIP-seq results.[32,41]

Considering the focus of this paper, we present CLIP-based approaches in greater detail. CLIP has been widely used to identify the RBP–RNA interaction regions, but this technique alone yields a reduced set of sequences containing the binding regions.[43] In practice, CLIP techniques use UV irradiation to covalent crosslink the RBP–RNA interaction; consequently the investigated protein is immunoprecipitated to isolate the complex, and partial RNase digestion of the bound transcript is used to select a short region of RNA attached to the protein. Nevertheless, only the joint action of CLIP with NGS makes it possible to obtain a transcriptome-wide set of interaction regions.[44] However, there is limited understanding of the crosslinking specificity at a physical level.[1] In order to overcome this limitation, ie, to identify in detail the crosslinking site and to improve the signal to noise ratio, several protocols have been proposed to determine the crosslinking sites: iCLIP,[45] PAR-CLIP[46] and HITS-CLIP.[47–49] In particular, the reverse transcription frequently stops at the crosslinking site in the iCLIP protocol. In HITS-CLIP, a nucleotide deletion is frequently found at the exact crosslinked amino-acid.[47] Alternatively, the PAR-CLIP protocol introduces an experimental variation at the beginning of the procedure, in order to facilitate the recognition of the interaction sites.[50] The nascent transcripts are labeled with a photo-reactive nucleoside (4-thiouridine) to print signatures inside and in the vicinity of the crosslinking site. In PAR-CLIP, the thymidine (T) to cytidine (C) transition (if nucleoside used is 4-thiouridine) near the crosslinking site is frequently found.

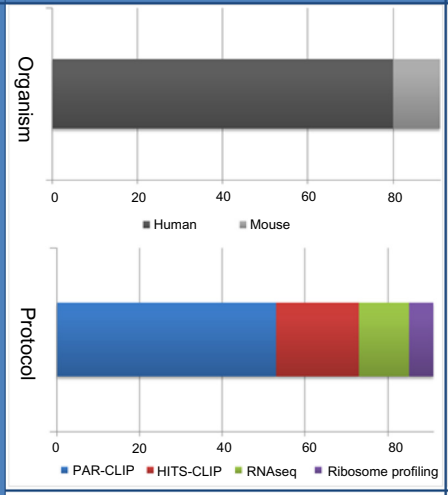
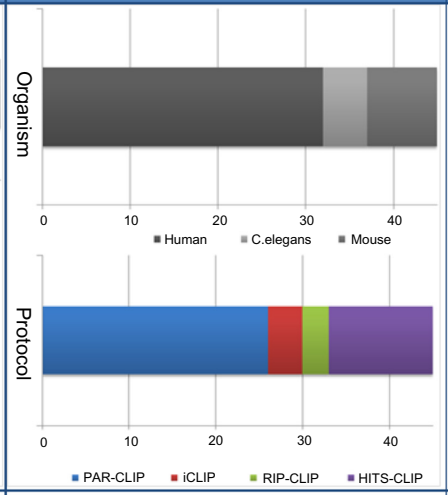
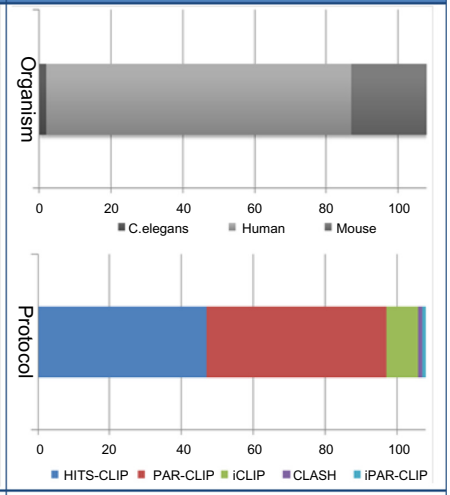
However, CLIP-seq experimental data need to be processed for further analysis. Therefore, computational methods for CLIP-seq data processing are necessary.

**Databases with CLIP-based data.** We present CLIP-seq data repositories that are publicly and freely available. This information is helpful to check whether there is another CLIP-seq dataset for the same protein under study or protein family. In addition, this information is quite valuable to design, test and validate new computational proposals.

Data sets obtained with CLIP-based protocols are frequently uploaded to public databases such as the Gene Expression Omnibus (GEO from NCBI)[54] and ArrayExpress (from EBI).[51] The authors of CLIP-based studies upload experimental data sets to share obtained results. In particular, the uploaded data often contain raw and processed data and a brief description of the data set. GEO and ArrayExpress are international public repositories that store high-throughput data obtained by the research community; the data sets uploaded in these two databases are publicly and freely available.

Recently, three databases have been developed specifically to store CLIP-based data: CLIPZ,[52] doRiNA,[53] and starbase v2.0.[55] These databases store CLIP-seq data, either uploaded directly by researchers or data sets available in the repositories previously mentioned. The data sets stored are carefully revised and processed. In addition, these databases provide additional useful functionalities for researchers in the field. In Table 1, we present the principal characteristics associated with the three databases. In particular, we present the data stored, the total number of data sets, and a graph with

**Table 1.** Databases with CLIP-seq data and associated characteristics.

| Database | CLIPZ | doRINA | StarBase V2.0 |
|---|---|---|---|
| Datasets |  |  |  |
| | 91 datasets | 45 datasets from 17 independent studies | 108 datasets from 37 independent studies |
| Website | http://www.clipz.unibas.ch | http://dorina.mdc-berlin.de/rbp_browser/dorina.html | http://starbase.sysu.edu.cn/index.php |
| Advantages | (1) Details of crosslinked nucleotide, adapter, sequencing technology<br>(2) Statistical summaries for RBP sites including annotations<br>(3) Cluster can be visualized at the transcript and genome level. | (1) A copy of the UCSC genome browser in local<br>(2) Predicts targets for a set or regulators (miRNA or RBP)<br>(3) Post-transcriptional regulation from the target or regulator point of view | (1) Annotations<br>(2) Associations to cancer statistically assessed<br>(3) Frequently updated |
| Reference | [52] | [53] | [55] |

the organisms present and the protocols used. Moreover, we provide a summary of the main advantages associated to each database, the website link, and reference.

CLIPZ is the first database published, specifically designed for CLIP-based data.[52] The database contains data sets published from 2010 to 2013. Most of the data available are CLIP-seq data, however, the database contains RNA-seq data from the same samples and Ribo-some profiling data. The CLIPZ database provides details for further processing such as the crosslinked nucleotide, the adapter, and the sequencing technology. In addition, the database provides statistical summaries, such as region preference, annotation summary, mutation plots, read quality, and read clusters length. The statistical summaries are presented for each data set or simultaneously for several selected data sets, which is particularly useful to make comparisons among several datasets.

doRiNA[53] is the second database published. This database contains data from 2010 to 2012, mainly CLIP-based data, but it also contains a few RIP-CHIP data sets. doRiNA has a local copy of the UCSC genome browser, which makes possible to have access to UCSC tracks. It is worth noting that doRiNA gives a post-transcriptional regulation view from the target or the regulator (miRNAs or RBP) point of view.

The third database is starbase v2.0.[55] This database contains data from 2010 to 2013. Initially, starbase was designed for microRNAs but the new version contains CLIP-seq data for a variety of RBPs. The database provides annotations for RBP sites, in particular lnc-RNA, mRNA, pseudogenes, and sncRNA. Moreover, it shows RBPs possible associations to cancer, which are statistically significant.

These freely available databases provide access to CLIP-based datasets. However, it is necessary to process the available data.

## Computational Pipeline to Process CLIP-Seq Data

Even though there is great room for further improvements in CLIP-seq data computational processing, several bioinformatic approaches have been proposed so far. The current

proposals address several steps in CLIP-seq data processing. In Figure 2, we present a pipeline with the most important steps in CLIP-seq data processing, and we associate computational tools designed for each step.

Here, we summarize these computational approaches for processing CLIP-seq data. We divided the computational approaches into categories depending on the scope. Moreover, in Table 2, we present additional characteristics for computational approaches specifically designed for CLIP-seq data.

**Read mapping and cluster detection.** The first step in CLIP-seq data processing is to map all the reads to the genome and transcriptome. During this step, at least one mismatch should be allowed because the experimental protocols induce nucleotide transitions (also known as mutations). Usually, the most frequently used algorithms to perform this step are Bowtie,[56] RMAP,[57] and Novoalign.[58] However, TopHat[59,60] is commonly used at this step, to identify exon–exon junctions.

Once the sequence reads are aligned to the genome and transcriptome, the following step is cluster detection. A cluster of reads is a group of reads, where a read belongs to a cluster if it overlaps at least one nucleotide to another read from the cluster. Several restrictions can be used to filter noise at this step. Usually, only reads with a length higher than a determined threshold are considered. In addition, clusters with a minimum number of unique reads are selected for binding site detection.

**Binding site detection.** After the cluster detection, the following step is reliable binding site detection. The main challenge at this step is to improve the signal to noise ratio, hence to remove background and false positives. The most common strategy to face this challenge is to analyze clusters distribution profiles. Computational approaches that use this strategy are WavClusteR,[61] PARalyzer,[62] Piranha,[63] PIPE-CLIP,[64] and dCLIP.[65] However, considering RNA structural features is also a good strategy that is present in GraphProt.[66] In this section, we briefly describe the above-mentioned computational proposals and present specific advantages.

It is worth noting that at this step it is definitely a plus to consider the number of sequences aligned to a specific cluster because this number strongly depends on the transcript abundance and cluster length.[63,65]

- PARalyzer[62] is the first computational approach designed for RBP site detection. This tool uses a non-parametric kernel density estimate and a classifier; it identifies the RBP sites based on a combination of T to C mutations and read density. PARalyzer can improve binding site recognition in data sets published.
- WavClusteR[61] is a computational tool proposed to overcome two problems in PAR-CLIP data processing. The first problem is the number of false positives, and the second is to improve cluster detection. Mutations present in experimental data are experimentally induced and also non-experimentally induced. In fact, nucleotide mutations are induced by the experimental protocol but in addition, several other factors cause mutations as well, such as sequencing errors, contamination with external RNA, and single-nucleotide polymorphisms (SNPs). WavClusteR uses a non-parametric two-component mixture model to distinguish experimentally from non-experimentally induced mutations, thus reducing the presence of false positives. In addition, the second part of wavClusteR exploits geometric

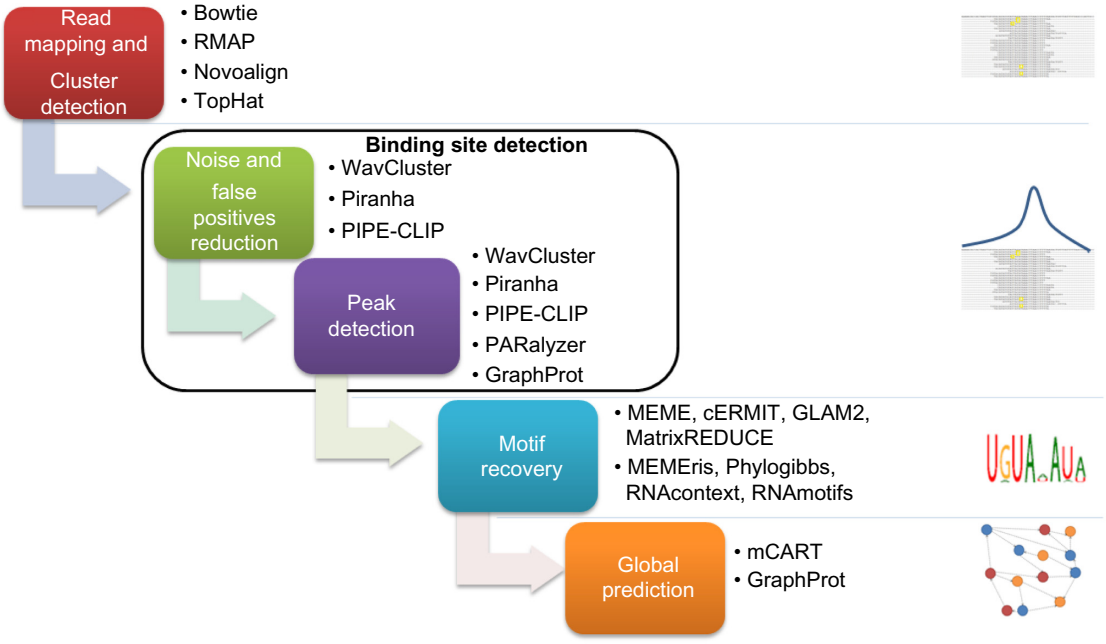



**Figure 2.** Steps for CLIP-seq data processing.

**Table 2.** Computational proposals specifically designed for CLIP-seq data processing.

| TOOL | YEAR | EXPERIMENTAL DATA USED | FOCUS | MAIN ADVANTAGE | RECOMMENDED CASE | AVAILABILITY | PROGRAMMING LANGUAGE |
|---|---|---|---|---|---|---|---|
| Paralyzer | 2011 | PAR-CLIP | Peak detection | Exploits T to C mutations to Improve Signal to noise ratio | PAR-CLIP data | www.genome.duke.edu/labs/ohler/research/PARalyzer/ | R |
| wavClusteR | 2012 | PAR-CLIP (BAM format) | Noise and false positives reduction Peak detection | Distinguishes between non-experimentally and experimentally induced transitions | PAR-CLIP data | https://github.com/FedericoComocglio/wavClusteR | R |
| Piranha | 2012 | CLIP-seq and RIP-seq (BED or BAM) | Noise and false positives reduction Peak detection CLIP-seq data comparison [correction for transcript abundance] | Corrects the reads dependence on transcript abundance | CLIP-seq and Transcript abundance data | http://smithlab.use.edu | Python |
| mCarts | 2013 | CLIP-seq | Sites prediction on different samples | Considers accessibility in local RNA secondary structures and cross-species conservation | RBP motif | http://zhanglab.c2b2.columbia.edu/index.php/MCarts | Perl |
| dCLIP | 2014 | CLIP-seq | Peak detection CLIP-seq data comparison [correction for transcript abundance] | Detects differential binding regions in comparing two CLIP-seq experiments | several CLIP-seq datasets and Transcript abundance data | http://qbrc.swmed.edu/software/ | Perl |
| PIPE-CLIP | 2014 | CLIP-seq (SAM or BAM) | Noise and false positives reduction Statistical assessment Peak detection | Provides a significance level for each identified candidate binding site | HITS-CLIP. iCLIP | http://pipeclip.qbrc.org/ Source code: https://github.com/QBRC/PIPE-CLIP | Python website available |
| GraphProt | 2014 | CLIP-seq and RNAcompete | Peak detection Sites prediction on different samples | Detects RBP motif secondary structure common characteristics. It estimates binding affinities | RBP motifs that are NOT located within single-stranded regions | http://www.bioinf_uni-freiburg.de/Software/GraphProt/ | Perl |

properties of the coverage function to identify reliable binding sites.

- Piranha[63] is a computational tool designed for site identification (peak calling), in CLIP-seq (HITS-CLIP, PAR-CLIP, iCLIP) and RIP-seq data. Piranha deals with three key challenges on computational site identification: (1) presence of noise and false positives, (2) resultant reads depend on transcript abundance, and (3) it is important to integrate different sources of information to improve peak calling. Piranha[63] uses a zero-truncated negative binomial distribution to model read counts, when additional information is available (covariates such as the transcript abundance), Piranha uses a zero-truncated negative binomial regression model. In addition, Piranha can compare CLIP-seq data from different samples because it corrects the reads dependence on transcript abundance.
- PIPE-CLIP[64] is a pipeline to identify binding regions. In PIPE-CLIP, the data are pre-processed to remove noise such as the PCR duplicates. Consequently, PIPE-CLIP identifies enriched clusters (considering cluster length effect on the number of reads) and reliable mutations. Each enriched cluster with at least one reliable mutation is selected as an RBP binding site.
- dCLIP[65] is a computational approach designed for quantitative CLIP-seq comparative analysis. dCLIP has two parts: normalization and RBP sites detection for comparison. The normalization step is necessary for an unbiased comparison. The second part is necessary to detect common or different sites for different CLIP-seq samples in order to perform a comparison.
- GraphProt[66] is a machine learning approach designed to identify RBP binding sites. This approach uses a training set to learn RBP binding preferences from high-throughput experimental data such as CLIP-seq and RNAcompete.[67] It uses a graph-kernel strategy to obtain a large set of features from the training set and any input data set. It should be noted that the features concern RNA sequence and also structure characteristics. GraphProt uses a support vector machine (SVM) to identify RBP sites using the set of features extracted. Moreover, when affinity data are available, GraphProt uses a support vector regression (SVR) to estimate affinities.

**Motif recovery.** The next step is to search the specific motif recognized by the RBP, among reliable binding sites. So far, two strategies are used for this purpose. The first one consists in using tools developed to detect motifs in DNA that consider only sequence information. The most frequently used tools for this strategy are MEME,[68] cERMIT,[69] GLAM2,[70] and MatrixREDUCE.[71] The second strategy consists in using motif recognition algorithms that integrate additional information to guide the motif search. Examples from the second strategy are MEMEris,[72,73] PhyloGibbs,[74,75] RNAcontext,[76,77] and RNAmotifs.[78]

It is worth noting that the second strategy permits to consider RNA-specific characteristics. In fact, MEMEris[72] uses RNA secondary structure to guide the motif search toward single-stranded regions, and PhyloGibbs[74] integrates conservation information. Moreover, RNAcontext[77] works on large-scale RNA-binding affinity datasets and provides the RNA motif in terms of sequence and structure. Finally, RNAmotifs[78] identifies multivalent regulatory motifs.

**Global prediction.** Once the RBP has a defined motif, we can use the motif to predict binding sites in a determined species. For this purpose, we should analyze motif occurrence characteristics[79] and predict candidate RBP binding sites. mCarts[80] and GraphProt[66] are two approaches proposed for this purpose.

In particular, mCarts is an algorithm based on a hidden Markov model that predicts functional RBP binding sites based on the number and spacing of motif sites, accessibility (RNA secondary structure) and conservation information. On the other hand, GraphProt is a machine learning approach that predicts candidate RBP binding sites within the same organism (training set data). In Table 2, we present additional characteristics such as availability.

**Considerations to reduce false positives.** In this section, we present two studies on CLIP-seq data, the results can be added to computational tools to reduce the false positives and improve the performance.

As already mentioned, in CLIP-based protocols covalent bonds are induced through UV crosslinking, only the RNA sites with strong bonds are selected through stringent washes. A study on PAR-CLIP background is presented in Ref. 81. This study presents possible sources of false positives such as RNAs bound to proteins different from the RBP of interest or false crosslinking events. Moreover, it shows that quantifying and taking into consideration possible sources of false positives are important to improve the recognition of the site specificity. As a result of the study, a set of background binding events in PAR-CLIP data is publicly available in GEO (GSE50989).

In addition, CapR is a tool designed to obtain a structural profile, which has been applied to CLIP-seq data. CapR[82] obtains a probability for each RNA base position, which reflects the location at determined structural contexts (CapR defines six contexts). Using these probabilities is possible to obtain a structural profile for an RNA sequence. Researchers can apply CapR on CLIP-seq data, so far, the obtained results are encouraging.

## Further Considerations

As indicated above, computational approaches are key to process CLIP-seq data and this field is expanding. In the RBP-RNA action, not only the sequence is important but also the RNA secondary structure. RBPs recognize the motif sequence content as well as the motif secondary structure. Even though

RNA secondary structure is considered in a few tools, it is a must and not just a plus.

Moreover, additional improvements can be achieved by integrating information about RBP domains, such as the one present in RBPDB database.[83]

In addition, advances on experimental field have provided techniques such as RNAcompete.[67] RNAcompete provides an affinity measure, independent on transcript abundance, it is an in vitro method. However, there is not a computational proposal that integrates information from both CLIP-seq and RNAcompete data, simultaneously.

Finally, an additional step can be added to the pipeline in Figure 2. After global prediction, integrative approaches for network inference can be used. This step is necessary to have a complete understanding of the specific role of RBPs.

## Author Contributions

PHRH conceived the idea and wrote the manuscript. PHRH and EF jointly developed the structure and arguments for the paper. EF made critical revisions. Both authors reviewed and approved of the final manuscript.

## Supplementary File

**Supplementary Figure S1.** Timeline for research on RNA binding proteins, this figure shows the number of publications on RNA binding proteins reported on PubMed since 1972.

## REFERENCES


1. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*. 2010;11(1):75–87.
2. Blackinton JG, Keene JD. Post-transcriptional RNA regulons affecting cell cycle and proliferation. *Semin Cell Dev Biol*. 2014;34:44–54.
3. Doxakis E. RNA binding proteins: a common denominator of neuronal function and dysfunction. *Neurosci Bull*. 2014;30(4):610–26.
4. Kapeli K, Yeo GW. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. *Front Neurosci*. 2012;6:144.
5. Kearse MG, Todd PK. Repeat-associated non-aug translation and its impact in neurodegenerative disease. *Neurotherapeutics*. 2014.
6. Lukong KE, Chang KW, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. *Trends Genetics*. 2008;24(8):416–25.
7. Kim MY, Hur J, Jeong S. Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB Rep*. 2009;42(3):125–30.
8. Ray D, Kazan H, Cook KB, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*. 2013;499(7457):172–7.
9. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage t4 DNA polymerase. *Science*. 1990;249(4968):505–10.
10. Trifllis P, Day N, Kiledjian M. Finding the right RNA: identification of cellular mRNA substrates for RNA-binding proteins. *RNA*. 1999;5(8):1071–82.
11. Ule J, Jensen K, Mele A, Darnell RB. Clip: a method for identifying protein-RNA interaction sites in living cells. *Methods*. 2005;37(4):376–86.
12. Änkö ML, Neugebauer KM. RNA–protein interactions in vivo: global gets specific. *Trends Biochem Sci*. 2012;37(7):255–62.
13. Larson DE, Sells BH. The function of proteins that interact with mRNA. *Mol Cell Biochem*. 1987;74(1):5–15.
14. Standart N, Jackson RJ. Regulation of translation by specific protein/mRNA interactions. *Biochimie*. 1994;76(9):867–79.
15. Siomi H, Dreyfuss G. RNA-binding proteins as regulators of gene expression. *Curr Opin Genet Dev*. 1997;7(3):345–53.
16. Darnell RB. Developing global insight into RNA regulation. *Cold Spring Harb Symp Quant Biol*. 2006;71:321–7.
17. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*. 2008;6(10):e255.
18. Mata J, Marguerat S, Bähler J. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends Biochem Sci*. 2005;30(9):506–14.
19. Halbeisen RE, Galgano A, Scherrer T, Gerber AP. Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell Mol Life Sci*. 2008;65(5):798–813.
20. Gerstberger S, Hafner M, Tuschl T. Learning the language of post-transcriptional gene regulation. *Genome Biol*. 2013;14(8):130.
21. Fedoroff NV. RNA-binding proteins in plants: the tip of an iceberg? *Curr Opin Plant Biol*. 2002;5(5):452–9.
22. Lorković ZJ. Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends Plant Sci*. 2009;14(4):229–36.
23. Tamburino AM, Ryder SP, Walhout AJ. A compendium of *Caenorhabditis elegans* RNA binding proteins predicts extensive regulation at multiple levels. *G3 (Bethesda)*. 2013;3(2):297–304.
24. Mattaj IW. A selective review of RNA–protein interactions in eukaryotes. *Mol Biol Rep*. 1990;14(2–3):151–5.
25. Richard S. Reaching for the stars: linking RNA binding proteins to diseases. *Adv Exp Med Biol*. 2010;693:142–57.
26. Pérez-Cañadillas JMP, Varani G. Recent advances in RNA–protein recognition. *Curr Opin Struct Biol*. 2001;ll(l):53–8.
27. Reyes-Herrera PH, Ficarra E. One decade of development and evolution of microRNA target prediction algorithms. *Genomics Proteomics Bioinformatics*. 2012;10(5):254–63.
28. Ascano M, Gerstberger S, Tuschl T. Multi-disciplinary methods to define RNA–protein interactions and regulatory networks. *Curr Opin Genet Dev*. 2013;23(1):20–8.
29. Reymond Sutandy FX, Hsiao FS, Chen CS. High throughput platform to explore RNA–protein interactomes. *Crit Rev Biotechnol*. 2014;0(0):1–9.
30. Gilbert C, Svejstrap JQ. RNA immunoprecipitation for determining RNA–protein associations in vivo. *Curr Protoc Mol Biol*. 2006;(27.4).
31. Milek M, Wyler E, Landthaler M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. *Semin Cell Dev Biol*. 2011;23(2):206–12.
32. König J, Zarnack K, Luscombe NM, Ule J. Protein–RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*. 2011;13(2):77–83.
33. George AD, Tenenbaum SA. Informatic resources for identifying and annotating structural RNA motifs. *Mol Biotechnol*. 2009;41(2):180–93.
34. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci U S A*. 2008;105(39):14885–90.
35. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*. 2010;16(6):1096–107.
36. Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through RNA structure. *Nat Rev Genet*. 2011;12(9):641–55.
37. Leontis NB, Westhof E. Analysis of RNA motifs. *Curr Opin Struct Biol*. 2003;13(3):300–8.
38. Hackermüller J, Meisner N-C, Auer M, Jaritz M, Stadler PF. The effect of RNA secondary structures on RNA–ligand binding and the modifier RNA mechanism: a quantitative model. *Gene*. 2005;345(1):3–12.
39. Iwakiri J, Kameda T, Asai K, Hamada M. Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. *Bioinformatics*. 2013;29(20):2524–8.
40. Kishore S, Luber S, Zavolan M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Brief Funct Genomics*. 2010;9(5–6):391–404.
41. Li X, Kazan H, Lipshitz HD, Morris QD. Finding the target sites of RNA-binding proteins. *Wiley Interdiscip Rev RNA*. 2013;5(1):111–30.
42. Zhao J, Ohsumi TK, Kung JT, et al. Genome-wide identification of polycomb-associated RNAs by rip-seq. *Mol Cell*. 2010;40(6):939–53.
43. Pellé R, Murphy NB. In vivo UV-cross-linking hybridization: a powerful technique for isolating RNA binding proteins, application to trypanosome miniexon derived RNA. *Nucleic Acids Res*. 1993;21(10):2453–8.
44. Licatalosi DD, Mele A, Fak JJ, et al. Hits-clip yields genome-wide insights into brain alternative RNA processing. *Nature*. 2008;456(7221):464–69.
45. Konig J, Zarnack K, Rot G, et al. iCLIP-transcriptome-wide mapping of protein–RNA interactions with individual nucleotide resolution. *J Vis Exp*. 2011;50:1–9.
46. Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by par-clip. *Cell*. 2010;141(1):129–41.
47. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from hits-clip data. *Nat Biotechnol*. 2011;29(7):607–14.
48. Darnell RB. Hits-clip: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA*. 2010;1(2):266–86.
49. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*. 2009;460(7254):479–86.

50. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of clip methods for identifying binding sites of RNA-binding proteins. *Nat Methods*. 2011;8(7):559–64.
51. Rustici G, Kolesnikov N, Brandizi M, et al. Arrayexpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013;41(Database issue):D987–90.
52. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res*. 2011;39(Database issue):D245–52.
53. Anders G, Mackowiak SD, Jens M, et al. doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res*. 2012;40(Database issue):D180–6.
54. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res*. 2013;41(Database issue):D991–5.
55. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starbase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale clip-seq data. *Nucleic Acids Res*. 2013.
56. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3):R25.
57. Smith AD, Chung WY, Hodges E, et al. Updates to the rmap short-read mapping software. *Bioinformatics*. 2009;25(21):2841–2.
58. Novocraft.com: Novoalign short read mapper. http://www.novocraft.com/main/downloadpage.php
59. Trapnell C, Pachter L, Salzberg SL. Tophat: discovering splice junctions with RNA-seq. *Bioinformatics*. 2009;25(9):1105–11.
60. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.
61. Sievers C, Schlumpf T, Sawarkar R, Comoglio F, Paro R. Mixture models and wavelet transforms reveal high confidence RNA–protein interaction sites in mov10 par-clip data. *Nucleic Acids Res*. 2012;40(20):el60.
62. Corcoran DL, Georgiev S, Mukherjee N, et al. Paralyzer: definition of RNA binding sites from par-clip short-read sequence data. *Genome Biol*. 2011;12(8):R79.
63. Uren PJ, Bahrami-Samani E, Burns SC, et al. Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*. 2012;28(23):3013–20.
64. Chen B, Yun J, Kim MS, Mendell JT, Xie Y. PIPE-CLIP: a comprehensive online tool for clip-seq data analysis. *Genome Biol*. 2014;15(1):R18.
65. Wang T, Xie Y, Xiao G. dCLIP: a computational approach for comparative clip-seq analyses. *Genome Biol*. 2014;15(1):R11.
66. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*. 2014;15(1):R17.
67. Ray D, Kazan H, Chan ET, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*. 2009;27(7):667–70.
68. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 1994;2:28–36.
69. Georgiev S, Boyle AP, Jayasurya K, Ding X, Mukherjee S, Ohler U. Evidence-ranked motif identification. *Genome Biol*. 2010;11(2):R19.
70. Frith MC, Saunders NFW, Kobe B, Bailey TL. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol*. 2008;4(4):el000071.
71. Foat BC, Morozov AV, Bussemaker HJ. Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*. 2006;22(14):el41–9.
72. Hiller M, Pudimat R, Busch A, Backofen R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res*. 2006;34(17):e117.
73. Bailey TL, Bodén M, Whiting-ton T, Machanick P. The value of position-specific priors in motif discovery using meme. *BMC Bioinformatics*. 2010;11:179.
74. Siddharthan R, van Nimwegen E. Detecting regulatory sites using PhyloGibbs. *Methods Mol Biol*. 2007;395:381–402.
75. Siddharthan R, Siggia ED, van Nimwegen E. PhyloGibbs: a Gibbs sampling motif finder that incorporates phytogeny. *PLoS Comput Biol*. 2005;l(7):e67.
76. Kazan H, Morris Q. RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Res*. 2013.
77. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*. 2010;6:el000832.
78. Cereda M, Pozzoli U, Rot G, et al. RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biol*. 2014;15(1):R20.
79. Ule J, Stefani G, Mele A, et al. An RNA map predicting nova-dependent splicing regulation. *Nature*. 2006;444(7119):580–6.
80. Zhang C, Lee K-Y, Swan-son MS, Darnell RB. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic Acids Res*. 2013.
81. Friedersdorf MB, Keene JD. Advancing the functional utility of par-clip by quantifying background binding to mRNAs and lncRNAs. *Genome Biol*. 2014;15(1):R2.
82. Fukunaga T, Ozaki H, Terai G, Asai K, Iwasaki W, Kiryu H. CAPR: revealing structural specificities of RNA-binding protein target recognition using clip-seq data. *Genome Biol*. 2014;15(1):R16.
83. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res*. 2011;39(Database issue):D301–8.